doi:10.3969/j.issn.1672-5565.2015.01.10

生物信息学软件研究的可视化分析

种乐熹[1],胡德华[1,2*]
(1.中南大学湘雅医学院医药信息系,长沙 410013;
2.湖南省高等学校医学信息研究重点实验室,长沙 410013)

摘　要:以 Web of Science 数据库为数据来源,利用 CiteSpace 和 UCINET 软件对发表在 Nucleic Acids Research 期刊上有关生物信息学软件研究的文献做了可视化分析,揭示了该领域的研究力量、作者团队与高被引作者、知识基础、期刊分布、研究热点与前沿,为生物信息学软件的研究和发展提供必要的参考依据。

关键词:生物信息学软件;知识图谱;共被引分析;研究热点;研究前沿
中图分类号:TP391　　**文献标志码**:A　　**文章编号**:1672-5565(2015)-01-054-13

Visualizing analysis of bioinformatics software research

CHONG Lexi[1], HU Dehua[1,2*]
(1.*Department of medical Information*, *Xiangya School of Medicine*, *Central South University*, *Changsha* 410013, *China*
2.*Key Laboratory of Medical Information Research*, *College of Hunan Province*, *Central South University*, *Changsha* 410013, *China*)

Abstract: Taking the web of science database as the data source, we analyzed visually the literatures about bioinformatics software published in Nucleic Acids Research Journal through CiteSpace and UCINET, the research effort, author teams, highly cited authors, knowledge base, journal distribution, research focuses and fronts in this field are explored to provide necessarys references for bioinformatics software research and development.

Keywords: Bioinformatics; Software Knowledge; Mapping Co-citation Analysis; Research Focus; Research Fronts

随着人类基因组计划的实施,通过基因组测序、蛋白质序列测定和结构解析等实验,分子生物学家提供了大量的有关生物分子的原始数据,需要利用现代计算技术对这些原始数据进行收集、整理、管理以便于检索使用,因而出现了生物信息学[1]。生物信息学产生的原因之一是利用计算机技术揭示大量而复杂的生物数据,而揭示的途径之一便是生物信息学软件。常用的软件有 BLAST、Clustal(X/W)、ANTHEPROT、RasMol、PyMOL、Swiss-PdbViewer 等。

近年来,生物信息学软件的研究得到了广泛关注,但大都集中于对相关软件的综述与归类[2-6]或者单一软件的介绍与应用[7-12]。国内外尚未有学者对生物信息学软件领域研究的整体情况进行分析,本文利用 CiteSpace、UCINET 等软件对生物信息学软件领域的研究情况进行可视化分析,揭示其研究热点和发展趋势,为相关研究者掌握研究现状和选择研究方向提供参考。

1　数据来源与研究方法

以 Web of Science 数据库为来源,构造检索式:出版物名称 = Nucleic Acids Research;索引 = SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CCR-EXPANDED, IC;时间跨度=2004—2013 年;分年选取 Web Server Issue 这一期,共检出记录 1 201 条。检索时间为 2014 年 4 月 25 日。

Nucleic Acids Research 属于国际权威期刊,其 2012 年的影响因子为 8.278,5 年影响因子为 8.055,主要刊登物理、化学、生物化学和生物学中核酸和蛋白质参与核酸代谢及其相互作用的最前沿研究,在

收稿日期:2014-11-26;**修回日期**:2015-01-09.
基金项目:国家社会科学基金项目(11BTQ044)。
作者简介:种乐熹,男,硕士研究生,研究方向:生物信息学;E-mail:chong_lx@163.com.
*通信作者:胡德华,男,教授,博士生导师,研究方向:生物信息学;E-mail:hudehua2000@163.com.

生物信息学、生命科学、遗传学与生物信息学等领域具有很高的影响力。同时,自 2004 年起,Nucleic Acids Research 每年 7 月单独出版"Web Server Issue"一期,重点介绍生物信息学软件的研究进展及重大研究成果。考虑到该期刊的权威性及数据获取的便捷性,本文选取 Nucleic Acids Research 期刊为研究对象。

导出上述检索结果的文献题录,主要采用学者陈超美博士基于 Java 平台开发的信息可视化分析软件 CiteSpace 进行分析,该软件通过对文献信息的可视化,能够直观地反应学科领域的发展轨迹、知识基础、研究前沿与热点等。同时,辅助 UCINET 软件对作者合作团队进行分析。

2　文献分布

对近 10 年生物信息学软件领域的相关文献信息量进行统计(见图 1),发现生物信息学软件的研究在 2005 年达到最高值,为 160 篇,之后呈下降趋势,表明 2005 年后对生物信息学软件的研究热度有所下降。2008 年起文献量稍有回升,但在 2010 年开始下降,之后趋于缓和,表明近几年对于生物信息学软件的研究热度稍有不足,同时表明对于生物信息学软件的研究达到了稳定状态。

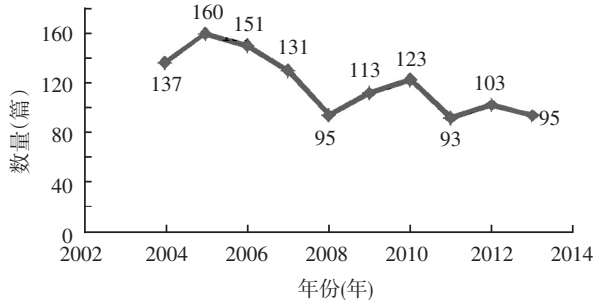

图 1　生物信息学软件研究的文献增长趋势

Fig.1　Literature distribution of bioinformatics software research

3　知识图谱可视化分析

3.1　研究力量分析

将数据导入到 CiteSpace 软件中,将 Node Types 设置为 Country 和 Institution,阈值设置为 T30,选择 Pathfinder 算法,其余选用默认值,得出生物信息学软件领域的研究力量分布图(见图 2)。其中圆形节点代表国家,发文量越多,节点越大,处于直线分支上的小节点代表机构。

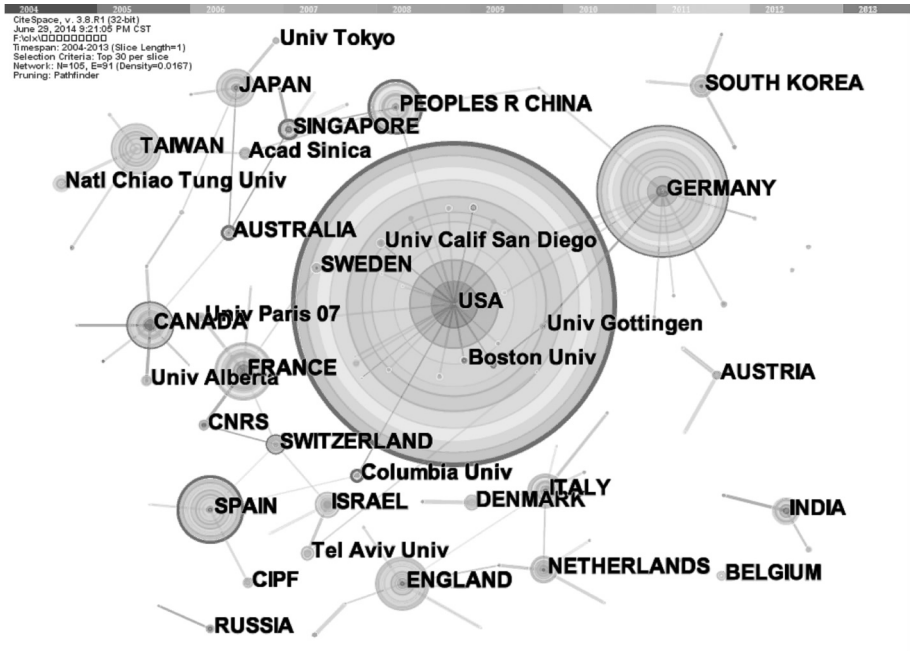

图 2　国家(地区)及机构分布知识图谱

Fig.2　Knowledge map of countries (regions) and institution distribution

不同国家和地区对于生物信息学软件的研究不尽相同,主要集中在美国、德国、西班牙、法国、英国、中国、加拿大等国家和地区,其中美国的发文量最多,为 340 篇,占到 10 年来总发文量的 28.3%;其次为德国,为 151 篇,占到 12.5%,西班牙紧随其后,为 72 篇,但在频次上远低于美国和德国,表明美国、德国、西班牙等国家在生物信息学软件方面的研究热度相对较高,其中以美国为盛。

从中心性看,美国的中心性最高,为 0.52。无论是发文频次还是中心性,美国都居于首位,表明美国对于生物信息学软件的研究热度和质量都较高;其次为中国大陆,为 0.37。新加坡(0.27)居于第三,随后是澳大利亚(0.20)。中国大陆的文献量虽与美国、德国等国家相比相对较少,但中心性位于前列,表明中国大陆对生物信息学软件的研究质量较高,具有较高的影响力,新加坡、澳大利亚等也具有一定的影响力。

生物信息学软件的研究机构主要集中在高校和研究院等机构,机构发文量相对于各国家而言,产出较少,其中台湾交通大学的发文量达到了 25 篇,以色列特拉维夫大学有 22 篇,台湾中央研究院为 20 篇,加拿大阿尔伯塔大学和法国国家科学研究院均为 18 篇;从中心性看,美国哥伦比亚大学的中心性较高,为 0.17,其次为德国的哥廷根大学(0.10)、台湾中央研究院(0.09)、特拉维夫大学(0.04)。无论是发文量还是中心性,机构的研究力量相对薄弱,尚未有高产出和高影响力的机构出现。在发文量排名前 15 的研究机构中,美国占了 5 席,中国台湾占了 3 席,表明二者的机构力量较强。

从国家(或地区)、研究机构的相互合作来看,总体合作状况并不密切,内部合作相比外部合作要多,某种程度上符合马太效应的特点,部分国家或地区的研究文献较多,且开展相互合作研究;部分国家或地区的相关研究较少,很少开展相互合作。

3.2 作者分析

通过对作者的发文量进行分析可以识别出某研究领域的高产作者。本研究共 1 201 篇文献,通过统计,选取发文量 10 篇以上的作者为高产作者,依次为 Dopazo J、Wishart DS、Morgenstern B、Tuffery P、Medina I、Alshahrour F、Nussinov R、Wolfson HJ、Tarrage J、Clote P,其发文量分别为 24、18、14、14、14、13、11、10、10、10 篇,表明这些作者在生物信息学软件领域研究中的贡献较高。

3.2.1 作者合作团队分析

分析科学研究中的作者合作关系对于资源共享、思想交流、知识传播、信息获取等方面都具有重要的意义[13]。本文选取发文量为 4 篇以上的作者,利用 UCINET 软件对生物信息学软件领域的作者合作关系进行分析。运行软件,得到核心作者的共现网络图谱。为形象地展现作者间的合作关系,剔除了孤立的节点,同时鉴于本文的合作网络较多,又剔除了若干个二人合作团队,得到作者合著图谱(见图 3),通过对其进行分析发现,生物信息学软件研究领域的核心作者合作团队较多,其中以 Dopazo J 为核心的团队最大,由 36 人构成;其次是以 Wishart DS 为核心的团队,由 25 人构成;以 Chen J、Zhang Y 等为核心的团队规模第三,由 17 人构成;同时,规模较大的还有一个 10 人团队、一个 9 人团队。

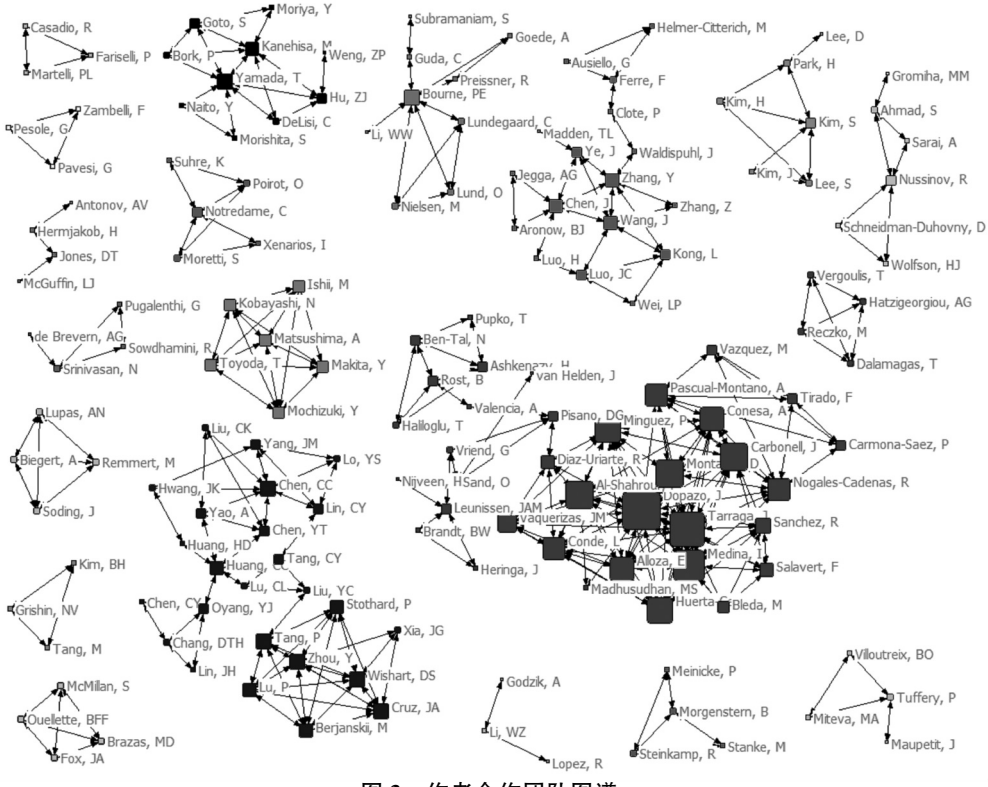

图 3 作者合作团队图谱

Fig.3 Map of co-author teams

规模最大的三个团队都研究和开发了大量的生物信息学软件,例如以 Dopazo J 为核心的团队开发了用于功能注释和基因组分析的 web 软件 BABELOMICS[14]、功能基因组数据分析工具 FatiGO +[15]、微阵列基因表达数据分析工具 GEPAS 等诸多软件[16];以 Wishart DS 为核心的团队开发了用于代谢组学数据分析和解释的工具 MetaboAnalyst[17]、质粒图谱的绘制和自动注释工具 PlasMapper[18]、细菌基因组自动注释工具 BASys 等[19];以 Chen J、Zhang Y 等为核心的团队开发了 GO 注释软件 WEGO[20]、蛋白质结构预测工具 SCRATCH[21]、miRNA 靶基因自动预测工具 miRU 等[22]。

图 3 中节点大小代表了作者的中心性,从图中可以看出以 Dopazo J 为核心的团队成员大都有着较高的中心性,表明这一团队的发文数量与质量均较高,是生物信息学软件研究领域的中坚力量;同时,以 Wishart DS 为核心的团队中的 Wishart DS、Cruz JA、Berjanskii M、Zhou Y、Lu P、Tang P 和以 Wishart DS 为核心的团队 Zhang Y、Chen J、Wang J 等作者也具有较高的中心性。进一步表明这三个团队对于生物信息学软件领域的研究具有较高的贡献。

以 Dopazo J 为核心的团队中心性最为突出,为进一步分析该团队的个体差异,单独对该团队进行分析(见图 4)。

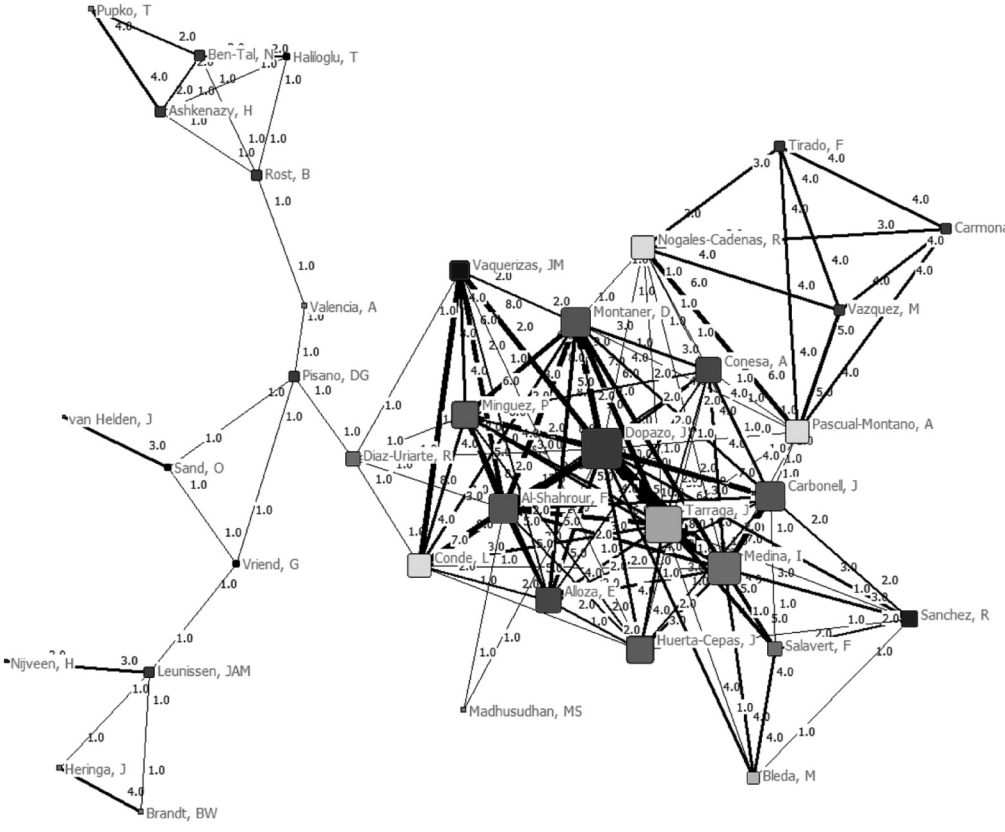

图 4　Dopazo J 团队分析图谱

Fig.4　Map of Dopazo J's team

图中两作者间连线的粗细表示其共现程度,连线上的数字表明其共现次数。从图中可以看出,Dopazo J 的中心性最高,与其他作者共现的次数也最多,表明其在生物信息学软件领域的研究较为活跃,成果最为突出。通过查询,发现其研究主要集中在生物信息学、系统生物学等领域,总计 h 指数为 55,i10 指数为 146。Tarrage J 的中心性和与其他作者共现的次数均位于第二,Medina I 的中心性和与其他作者共现的次数位于第三,Alshahrour F、Montaner D、Carbonell J 三人其次,表明这些作者在生物信息学软件研究的贡献较高,研究成果较为突出。进一步分析,发现该团队可从 Diaz-Uriarte R 划分为左右两部分,而左边各作者的中心性及与其他作者共现的次数相对右边各作者较少,表明 Diaz-Uriarte R 促进了作者之间的交流合作,但合作的力度并不强。

3.2.2　作者共被引分析

美国德雷克赛大学怀特(White)博士认为,作者共引频次越高则作者学术相关性越强[23]。将 Node Types 设置为 Cited Author,阈值设置为 T30,其余选用默认值,得到作者共被引图谱(见图 5),图中节点代表共被引作者,节点大小代表共被引频次。

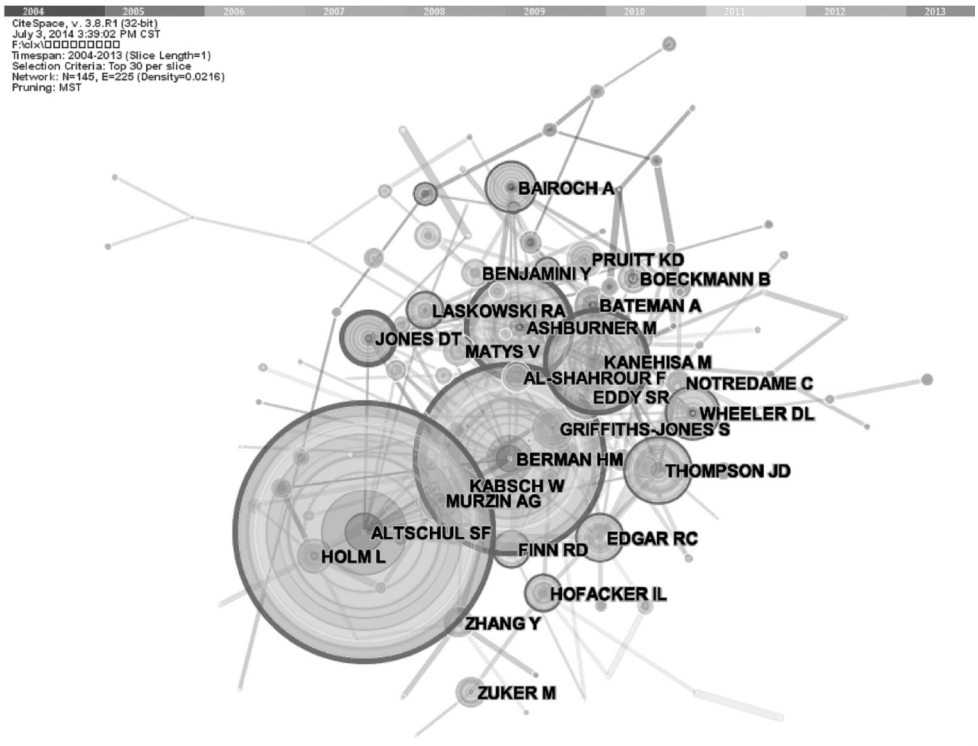

图 5 作者共被引知识图谱

Fig.5 Knowledge map of author co-citation

从图 5 中可以看出,共被引频次最高的是 ALTSCHUL SF(260 次)。ALTSCHUL SF 出生于 1957 年,是美国的一位数学家,现主要从事生物信息学的研究,设计的算法(Karlin-Altschul 算法及其继任算法)被广泛应用在生物信息学领域,发表的所有文献总被引频次为 122 642 次,h 指数为 44,i10 指数为 57。ALTSCHUL SF 等人于 1990 年提出 BLAST(Basic Local Alignment Search Tool)算法,被广泛使用在蛋白质 DNA 序列的分析问题中,在其他序列相似性比对中也有应用。

其次是 BERMAN HM(192 次),她出生于 1943 年,是 RCSB PHD 的负责人,同时也是 Nucleic Acid Database 的创始人,是一位结构生物学家,现主要从事结构生物学、结构生物信息学和数据库的研究,发表的所有文献总被引频次为 35 498 次,h 指数为 71,i10 指数为 168;随后是 ASHBURNER M(111 次),生于 1942 年,是英国遗传学、生物信息学领域的著名学者,发表过 Gene Ontology: tool for the unification of biology[24]等高被引文献;KANEHISA M 是日本京都大学生物信息学领域的著名学者;THOMPSON JD 是法国斯特拉斯堡大学生物学学者,发表过关于 CLUSTAL 软件[25-27]的高被引文献。JONES DT、EDGAR RC、WHEELER DL、BAIROCH A、GRIFFITHS-JONES S 等都是生物学等相关领域的著名学者,学术成果突出。

3.3 知识基础分析

通过对学科领域的文献信息可视化分析能够使研究者直观地辨识出学科前沿的演化路径及学科领域的经典基础文献[28]。将数据导入到 CiteSpace 软件中,将 Node Types 设置为 Cited Reference,阈值设置为 T30,其余选用默认值,得到共被引文献的可视化图谱,分别以 Timeline 和 Cluster 方式显示,得到奠基性文献和核心文献的知识图谱,图中圆形节点代表共被引文献,节点大小代表共被引频次。

对生物信息学软件研究的知识基础从两个方面进行分析,即生物信息学软件研究的早期奠基性文献、高被引与高中心性文献,它们构成了生物信息学软件研究的脉络,形成了坚固的基础。

3.3.1 奠基性文献分析

从图 6 中可以看出有 4 篇于 20 世纪 80 年代发表的早期文献(见表 1),第 1 篇是 Smith TF 和 Waterman MS 于 1981 年发表在 Journal of Molecular Biology 期刊上的 Identification of common molecular subsequences 一文,该文在 Google Scholar 中的被引频次为 7 704,在 Web of Science 中的被引频次为 3 741,在所选文献集中的被引频次为 34 次,中心性为 0.07。在文中,Smith 和 Waterman 提出了著名的公共子序列识别算法(Smith-Waterman 算法)[29],

Smith-Waterman 算法作为 Needleman-Wunsch 算法的改进,提供了很好的比对性能和更大的灵活性。

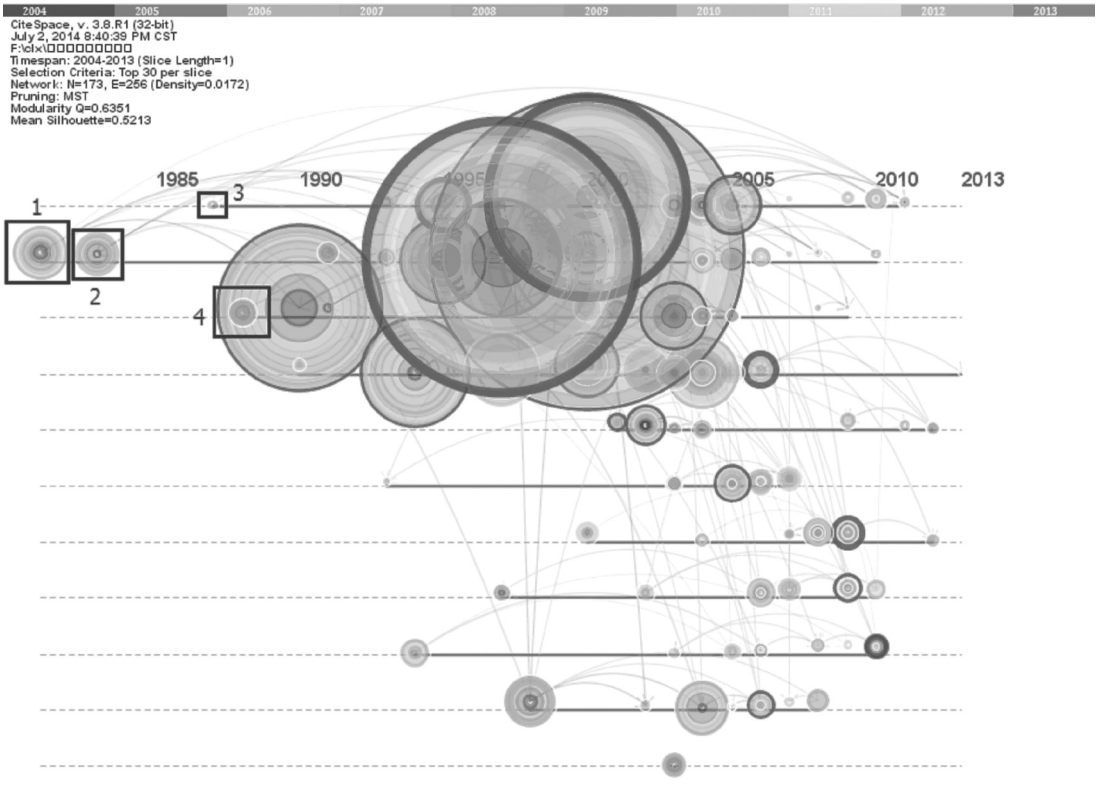

图 6　奠基性文献的时间序列图谱

Fig.6　Time line map of foundational literatures

第 2 篇是 Kabsch W 和 Sander C 于 1983 年发表在 Biopolymers 期刊上的 Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features 一文,在 Google Scholar 中的被引频次为 9 509 次,在 Web of Science 中的被引频次为 7 956 次,在所选文献集中的被引频次为 30 次。Wolfgang Kabsch 和 Chris Sander 在文中提出了应用于蛋白质结构中的氨基酸残基进行二级结构构像分类的标准算法[30]。

表 1　生物信息学软件研究的奠基性文献

Table 1　Foundational literatures of bioinformatics software research

年份	频次	中心性	被引文献
1981	34	0.07	SMITH TF, 1981, J MOL BIOL, V147, P195, DOI 10.1016/0022-2836(81)90087-5
1983	30	0.03	KABSCH W, 1983, BIOPOLYMERS, V22, P2577, DOI 10.1002/BIP.360221211
1987	13	-	SAITOU N, 1987, MOL BIOL EVOL, V4, P406
1988	20	0.01	PEARSON WR, 1988, P NATL ACAD SCI USA, V85, P2444, DOI 10.1073/PNAS.85.8.2444

第 3 篇是 Saitou N 和 Nei M 于 1987 年发表在 Molecular Biology and Evolution 上的 The neighbor-joining method: a new method for reconstructing phylogenetic trees 一文,在 Google Scholar 中的被引频次为 36 178,在 web of science 中的被引频次为 30 657 次,在所选文献集中的被引频次为 13 次。Saitou 和 Nei 在文中提出了用于构建系统发育树的邻接法(Neighbor-joining Method)[31],该方法通过确定距离最近或距离相邻的成对分类单位来使系统树的总长达到尽可能小。

第 4 篇是 Pearson WR 和 Lipman DJ 于 1988 年发表在 Proceedings of the National Academy of Sciences 上的 Improved tools for biological sequence comparison 一文,被引频次为 20,在 Google Scholar 中

的被引频次为 11 271。Pearson 和 Lipman 对 FASTP 算法进行了改进[32]，提出了序列比较算法(FASTA)。

3.3.2 核心文献分析

对于核心文献的分析主要从高被引文献(见表 2)和高中心性文献(见表 3)展开，从图 7 中可以看出被引频次最高的文献是 BERMAN HM 于 2000 年发表的 The Protein Data Bank 一文，BERMAN HM 在文中介绍了 PDB 数据库的目标[33]、数据存储和访问系统以及如何获取进一步信息，并对资源的发展做了短期规划。PDB 是一个专门收录蛋白质及核酸的三维结构资料的数据库，PDB 及其所提供的软件对公众免费使用。

表 2 生物信息学软件研究的高被引文献

Table 2 Highly cited literatures of bioinformatics software research

频次	年份	中心性	被引文献
171	2000	0.18	BERMAN HM, 2000, NUCLEIC ACIDS RES, V28, P235, DOI 10.1093/NAR/28.1.235
146	1997	0.35	ALTSCHUL SF, 1997, NUCLEIC ACIDS RES, V25, P3389, DOI 10.1093/NAR/25.17.3389
105	2000	0.36	ASHBURNER M, 2000, NAT GENET, V25, P25
91	1990	0.34	ALTSCHUL SF, 1990, J MOL BIOL, V215, P403, DOI 10.1006/JMBI.1990.9999
63	1994	0.13	THOMPSON JD, 1994, NUCLEIC ACIDS RES, V22, P4673, DOI 10.1093/NAR/22.22.4673
49	1995	0.1	MURZIN AG, 1995, J MOL BIOL, V247, P536, DOI 10.1016/S0022-2836(05)80134-2
45	2004	0.03	EDGAR RC, 2004, NUCLEIC ACIDS RES, V32, P1792, DOI 10.1093/NAR/GKH340
41	1997	0.04	ALTSCHUL SF, 1997, NUCLEIC ACIDS RES, V25, P3402

表 3 生物信息学软件研究的高中心性文献

Table 3 High central literatures of bioinformatics software research

频次	年份	中心性	被引文献
105	2000	0.36	ASHBURNER M, 2000, NAT GENET, V25, P25
146	1997	0.35	ALTSCHUL SF, 1997, NUCLEIC ACIDS RES, V25, P3389, DOI 10.1093/NAR/25.17.3389
91	1990	0.34	ALTSCHUL SF, 1990, J MOL BIOL, V215, P403, DOI 10.1006/JMBI.1990.9999
30	1995	0.27	BENJAMINI Y, 1995, J ROY STAT SOC B MET, V57, P289
171	2000	0.18	BERMAN HM, 2000, NUCLEIC ACIDS RES, V28, P235, DOI 10.1093/NAR/28.1.235
19	2006	0.18	HULL D, 2006, NUCLEIC ACIDS RES, V34, , DOI 10.1093/NAR/GKL320
36	2003	0.16	BOECKMANN B, 2003, NUCLEIC ACIDS RES, V31, P365, DOI10.1093/NAR/GKG095
31	2004	0.16	AL-SHAHROUR F, 2004, BIOINFORMATICS, V20, P578, DOI10.1093/BIOINFORMATICS/BTG455

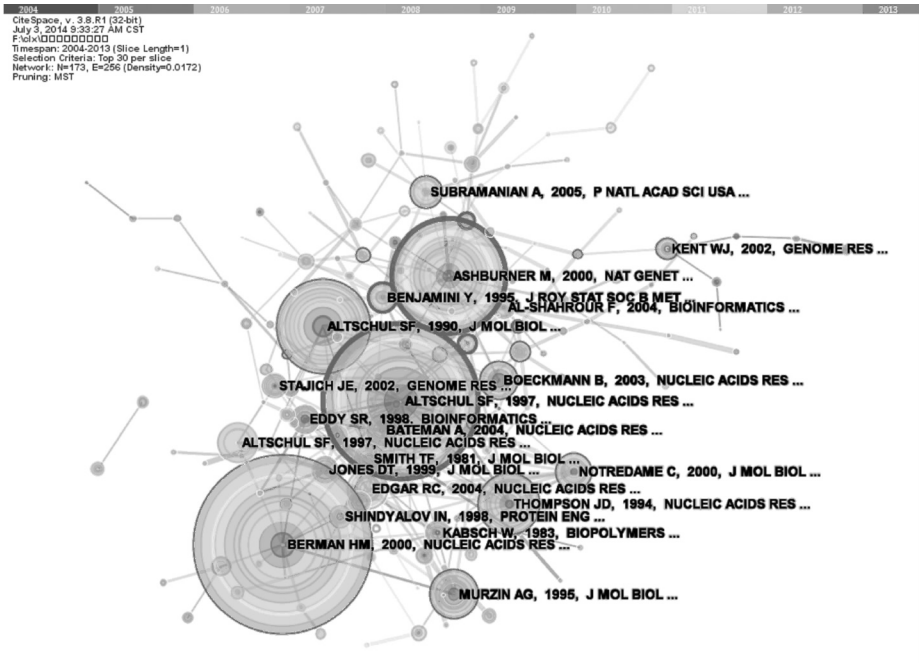

图 7 生物信息学软件研究的核心文献图谱

Fig.7 Knowledge map of core literatures

其次是 ALTSCHUL SF 于 1997 年发表的 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs 一文,文中介绍了 BLAST 程序的不足,并阐述了新的程序——PSI-BLAST[34]。

频次第 3 的是 ASHBURNER M 于 2000 年发表的 Gene Ontology: tool for the unification of biology 一文,文中介绍了 Gene Ontology[35],它由 3 部分组成,即分子功能(molecular function)、生物过程(biological process)、细胞组成(cellular component)。蛋白质或基因可以通过 ID 对应或者序列注释的方法找到与之对应的 GO 号,而 GO 号可对应到 Term,即功能类别或者细胞定位。

中心性排名前两位的文献分别是被引频次排名第 3 和第 2 的文献,表明这两篇文章在生物信息学软件领域具有一定的重要性。

中心性排名第 3 的 ALTSCHUL SF 于 1990 年发表的 Basic local alignment search tool 一文,文中提出了著名的 BLAST 算法[36],BLAST 算法是一种基于局部序列比对的序列比对算法,能够实现比较两段核酸或者蛋白序列之间的同源性的功能,它能够快速的找到两段序列之间的同源序列并对比对区域进行打分以确定同源性的高低。

值得注意的是排名第 4 的文献 Controlling the false discovery rate: a practical and powerful approach to multiple testing,虽然被引频次较低,但有着较高的中心性,表明该文献有着较高的价值,文中指出解决多样性问题的常规方法具有缺陷,并提出了一种不同的方法来解决多样性问题[37]。

3.4　期刊共被引分析

对某研究领域文献的来源期刊进行分析是掌握该研究领域核心期刊群的最有效方法,也是研究人员研究该领域的重要情报源[38]。将 Node Types 设置为 Cited Reference,阈值设置为 T30,其余选用默认值,得到期刊共被引知识图谱(见图 8)。

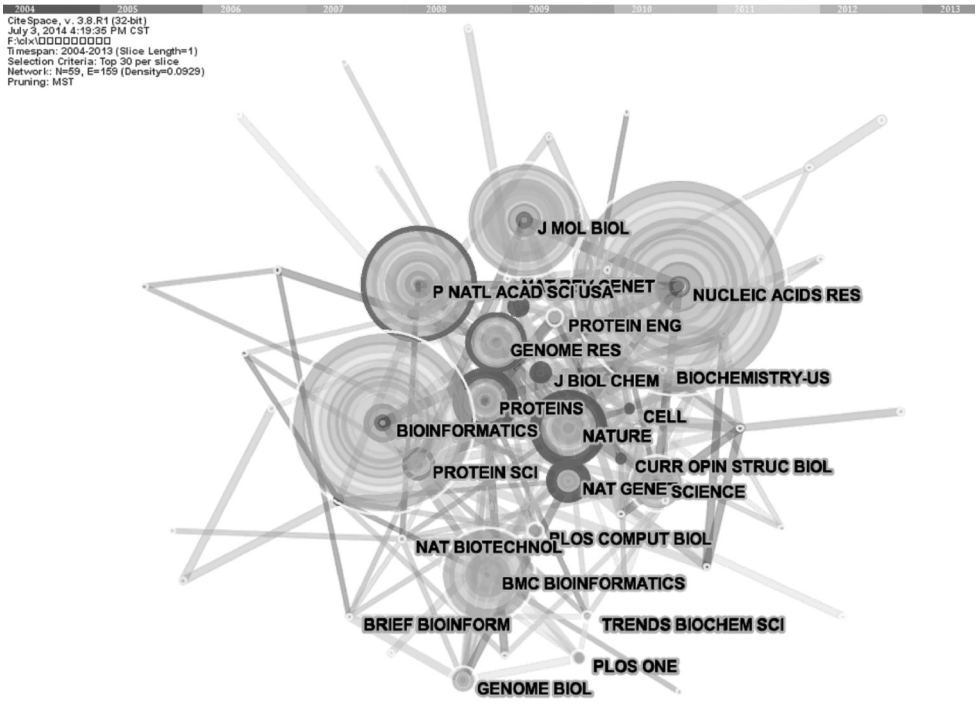

图 8　期刊共被引知识图谱

Fig.8　Knowledge map of journal co-citation

从共被引频次看,生物信息学软件研究领域的重要期刊有 NUCLEIC ACIDS RES、BIOINFORMATICS、J MOL BIOL、P NATL ACAD SCI USA、BMC BIOINFORMATICS、NATURE、SCIENCE、GENOME RES,它们的共被引频次依次为 1 107、902、603、559、525、396、366、362,这些期刊作为生物信息学研究领域的重要期刊,刊载了生物信息学软件研究领域的众多高质量文献,反映了生物信息学软件领域的重要研究内容。

从中心性看,生物信息学软件领域的重要期刊有 NATURE、NAT GENT、J BIOL CHEM、GENOME RES、PROTEINS、J AM CHEM SOC,它们的中心性分别为 0.36、0.24、0.22、0.20、0.20、0.20。

结合共被引频次与中心性进行分析,发现生物信息学软件研究领域的重要期刊主要来自 NUCLEIC

ACIDS RES、BIOINFORMATICS、J MOL BIOL、P NATL ACAD SCI USA、BMC BIOINFORMATICS 等生物学、生物信息学、化学等领域的期刊,也有 NATURE、SCIENCE 等自然科学类综合性期刊。被引期刊的质量都较高,例如 2012 年 NUCLEIC ACIDS RES 的影响因子为8.278, BIOINFORMATICS 为 5.323,J MOL BIOL 为 3.905,P NATL ACAD SCI USA 为 9.737,BMC BIOINFORMATICS 为 3.024,而 NATURE 和 SCIENCE 的影响因子更是高达 38.597和 37.027,这些高质量期刊在生物信息学软件的研究中起着举足轻重的作用。

3.5 研究热点分析

将 Node Types 设置为 Keyword,将阈值设置为 T30,选择 Pathfinder 算法,其余选择默认设置,运行软件,得到生物信息学软件研究热点知识图谱(见图 9)。圆圈的颜色和厚度分别表示关键词出现的时间和数量,每个节点分别表示一个关键词,节点的大小表示关键词出现的频次,节点之间的连线越粗表示关键词之间的共现越高。

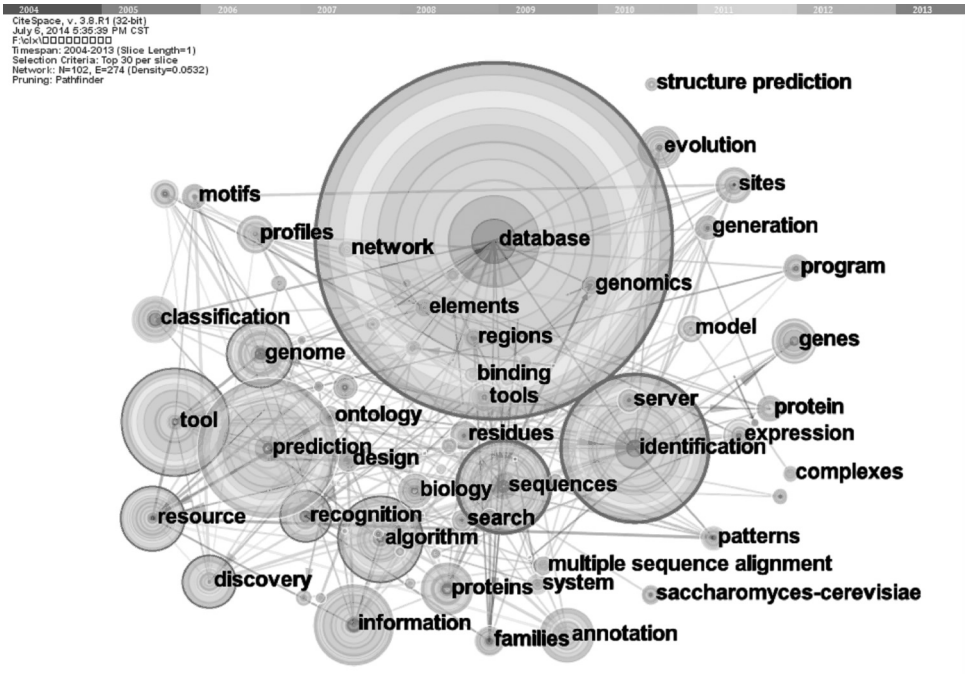

图 9 生物信息学软件研究热点知识图谱

Fig.9 Knowledge map of research focus

关键词是一篇文章的核心与精髓,是文章主题的高度概括及凝炼,频次高的关键词常被用来确定一个研究领域的热点问题[39]。频次最高的是 database,达到 320 次。生物信息学软件往往集成在相关生物信息学数据库中,比如 EBI 主页上的工具 Tools,NCBI 主页上的工具 Blast,因而对于软件的研究往往离不开数据库,database 表现出的高频次并不意外。对高频关键词进行分析(见表 4),发现生物信息学软件研究的热点主要集中在软件功能上,identification、prediction、discovery、annotation、recognition 等关键词表明研究热点可以划分为识别、预测、研发、注释等主题。同时,sequences、information、resource、proteins 等高频关键词表明研究热点主要集中在基因序列和蛋白质序列上,algorithm 表明相关算法的研究也是研究热点之一。

同时,中心性较高的关键词在整个网络中具有重要的连接作用,关键词的中心性越强,表明其控制的关键词之间的信息流越多,因而中心性高的关键词也能代表某领域的研究热点[40]。从表 5 中可以看出,高中心性的关键词都是高频关键词,database、identification、sequences、algorithm、genome 等关键词的中心性排名最高,体现出数据库、识别、(基因)、序列、算法等占有较高的研究热度。

通过对高频关键词进行聚类(见图 10),发现生物信息学软件领域的研究可以聚成 6 个大类,分别是 protein complexe、opposite gene expression signature、depth、visualization software、predicting subcellular localization、soba (sequence ontology bioinformatic analysis),这 6 个大类代表了生物信息学软件研究的热点,即蛋白质复合体、对生基因表达特征、生物信息学数据的深入分析、可视化软件、亚细胞定位预测、序列本体论生物信息学分析工具。

表 4　高频关键词表(≥50)

Table 4　High frequency keywords list(≥50)

序号	频次	关键词
1	320	database
2	134	identification
3	133	prediction
4	101	tool
5	86	sequences
6	81	algorithm
7	78	information
8	63	genome
9	63	resource
10	55	annotation
11	54	discovery
12	51	proteins
13	51	recognition

表 5　高中心性关键词表(≥0.1)

Table 5　High centrality keywords list(≥0.1)

序号	中心性	关键词
1	0.3	database
2	0.28	identification
3	0.2	sequences
4	0.15	algorithm
5	0.15	genome
6	0.14	tool
7	0.11	resource
8	0.11	discovery
9	0.11	recognition
10	0.1	families

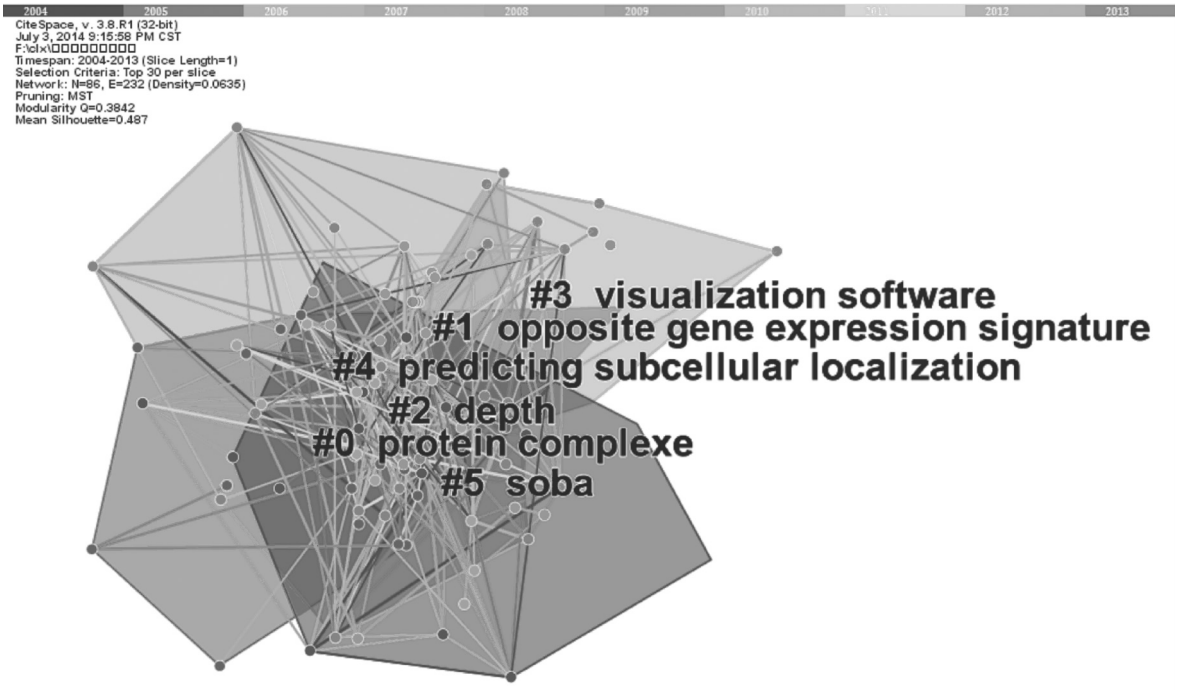

图 10　高频关键词聚类图谱

Fig.10　Knowledge map of clustering high frequency keywords

3.6　研究前沿分析

本文通过对生物信息学软件领域研究文献关键词的突变情况来确定其研究前沿,将网络节点类型设置为 Keyword,并将 Term Type 选定为 Burst Terms,选取 Top 50,其余采用默认值,运行 citespace,并以时区示图(Timezone)方式显示,得到研究前沿知识图谱(见图 11)。

在图中圆形节点代表关键词,圆形节点的大小表示关键词的频次。方形节点代表了突变词,方形节点的大小表示突变词的权重,其位置反映出该突变词出现的时间。对突变词按权重进行排序,权重最高的结果如表 6 所示。

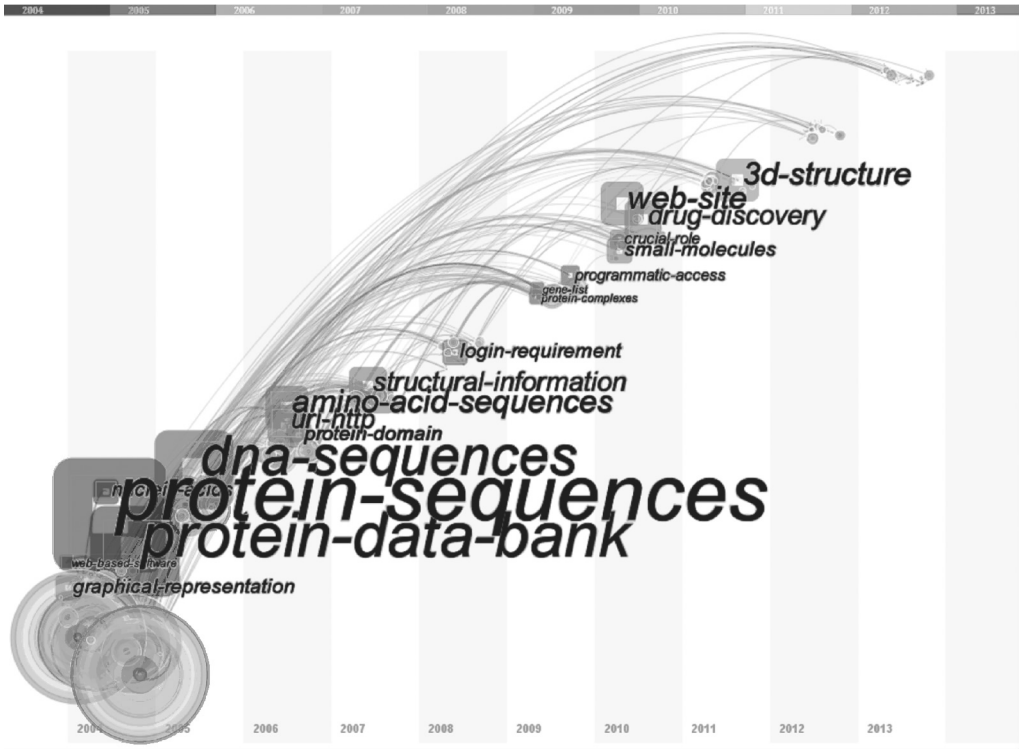

图 11 研究前沿知识图谱

Fig.11 Knowledge map of research fronts

表 6 生物信息学软件研究的突变词表

Table 6 Burst keywords of bioinformatics software research

突变词	权重	年份
protein-sequences	11.275 3	2004
protein-data-bank	8.887 5	2004
DNA-sequences	7.532 6	2005
3d-structure	4.251 2	2011
web-site	4.126 9	2010
amino-acid-sequences	4.080 9	2006
drug discovery	3.535 9	2010
URL-http	3.496 6	2006
structural information	3.503 1	2007
login requirement	2.428 6	2008

权重排名前三的突变词分别是 protein-sequences(蛋白质序列)、protein-data-bank(蛋白质数据库)、DNA-sequences(DNA 序列),表明生物信息学软件领域的研究前沿主要集中在蛋白质序列、DNA 序列的分析上,protein-data-bank 之所以出现较高的权重在于生物信息学相关软件越来越多的集成在相关数据库中。

其它突变词还包括 3d-structure(三维结构)、web-site(网站)、amino-acid-sequences(氨基酸序列)、drug discovery(药物研发)、URL-http(网址)、structural information(结构信息)、login requirement(登录要求)、gene-list(基因列表)、programmatic-access(编程访问)、crucial role(重要角色)、protein-complexes(蛋白质复合体)、small-molecules(小分子)、protein domain(蛋白质结构域)、web-based-software(web 软件)、nucleic-acids(核酸)、graphical-representation(图形化表示),这些都是生物信息学软件领域的研究前沿之一。通过对其进行分析,发现主要集中在软件功能与软件自身属性上,具体而言,解决三维结构、氨基酸序列、药物研发、结构信息、基因列表、蛋白质复合体、小分子、蛋白质结构域等问题是生物信息学软件领域的研究前沿之一。网站、网址、登录要求、web 软件等体现了软件属性是研究前沿之一。

通过对查阅近几年文献,发现了大量体现研究前沿的研究,如 ModeRNA: a tool for comparative modeling of RNA 3D structure[41]、《分子三维结构可视化软件建模的研究与实现》等文章体现了三维结构是研究前沿之一[42], The case for open-source software in drug discovery[43]、《生物信息学分析软件的开发及在药学领域的应用》等文章体现了药物研发是研究前沿之一[44]。同时,PLAAC: a web and command-line application to identify proteins with Prion-Like Amino Acid Composition[45]、SKINK: a web server for string kernel based kink prediction in α-helices 等文章中介绍了 PLAAC、SKINK 等软件的开发利用[46],包括了对软件属性和功能的描述,侧面

证明了本研究所探测的研究前沿的正确性。

4 结　论

通过对生物信息学软件领域的研究力量、知识基础、核心作者与期刊、研究热点和研究前沿的分析,对生物信息学软件领域的研究现状进行总结,得出如下结论。

(1)不同国家(地区)和机构对于生物信息学软件的研究不尽相同,主要集中在美国、德国、西班牙、法国、英国、中国台湾、中国大陆、加拿大等国家和地区,其中美国最为突出,中国大陆、新加坡和澳大利亚等国家的影响力也较大;同时,生物信息学软件的研究机构主要集中在高校和研究院等机构,但机构的研究力量相对薄弱,尚未有高产出和高影响力的机构出现。

(2)从两个角度分析了生物信息学软件研究的知识基础:一是奠基性文献,这组文献包括 Smith TF 和 Waterman MS、Kabsch W 和 Sander C、Saitou N 和 Nei M、Pearson WR 和 Lipman DJ 等发表的四篇文章,在文章中分别提出了 Smith-Waterman 算法、应用于蛋白质结构中的氨基酸残基进行二级结构构像分类的标准算法、邻接法、FASTA 算法,这些算法对相关软件的研究起着重要指导作用;二是高被引和高中心性文献,这组文献主题包括 PDB 数据库、PSI-BLAST、Gene Ontology、BLAST 算法及多样性问题解决方法,对生物信息学软件的研究起着重要的支柱作用。

(3)Dopazo J、Wishart DS、Morgenstern B 等作者构成了生物信息学软件领域的高产作者,为该领域的研究起到了较大的贡献。同时,挖掘出以 Dopazo J、Wishart DS、Chen J 和 Zhang Y 为核心的三个作者合作团队,其中以以 Dopazo J 为核心的团队中心性最为显著,团队成员在该领域的研究成果最为突出。在作者的共被引分析中,发现 ALTSCHUL SF、BERMAN HM、ASHBURNER M 等作者的被引频次最高,对该领域的研究起着重要的启发作用。

(4)对共被引期刊进行分析,发现 NUCLEIC ACIDS RES、BIOINFORMATICS、J MOL BIOL、P NATL ACAD SCI USA、BMC BIOINFORMATICS、NATURE、SCIENCE 等期刊在生物信息学软件的研究中起着举足轻重的作用。

(5)对研究热点与前沿进行分析发现,生物信息学软件的研究热点主要集中在对生物信息学相关软件功能的研究上,表现在识别、预测、研发、注释等主题的研究上,进一步分析,发现研究热点可以分为蛋白质复合体、对生基因表达特征、生物信息学数据的深入分析、可视化软件、亚细胞定位预测、序列本体论生物信息学分析工具等大类;解决三维结构、氨基酸序列、药物研发、结构信息、基因列表、蛋白质复合体、小分子、蛋白质结构域等问题和软件自身属性是研究前沿。

参考文献(References)

[1] 毛黎明. 分布式并行处理与复杂网络在蛋白质折叠中的应用[D]. 武汉:武汉理工大学, 2005.
MAO Liming. The application of distributed parallel processing and complex network in protein folding[D]. Wuhan: Wuhan University of Technology, 2005.

[2] GILBERT D D. Bioinformatics software resources [J]. Briefings in Bioinformatics, 2004, 5(3): 300-304.

[3] GILBERT D.Pise: Software for building bioinformatics webs [J].Briefings in Bioinformatics, 2002, 3(4): 405-409.

[4] KUMAR S, DUDLEY J. Bioinformatics software for biologists in the genomics era[J]. Bioinformatics, 2007, 23(14): 1713-1717.

[5] 牛钦王,陈建平. 生物信息学常用方法及其应用软件概述[J]. 热带医学杂志,2012,12(7):908-910.
NIU Qinwang, CHEN Jianping. Overview on common methods and application software of bioinformatics[J]. Journal of Tropical Medicine, 2012, 12(7): 908-910.

[6] 徐思敏. RNA 生物信息相关软件概述[J]. 科技信息, 2008, (14):66-67.
XU Simin. Overview on RNA bioinformatics related software[J].Scinece & Technology Information, 2008, (14): 66-67.

[7] GOTO N, PRINS P, NAKAO M, et al. BioRuby: bioinformatics software for the Ruby programming language [J].Bioinformatics, 2010, 26(20): 2617-2619.

[8] HUANG Y, WANG J Y, WEI X M, et al. Bioinfo-Kit: A Sharing Software Tool for Bioinformatics[J].Applied Mechanics and Materials, 2014, 472: 466-469.

[9] GOMEZ J, GARCIA L J, SALAZAR G A, et al. BioJS: an open source JavaScript framework for biological data visualization[J].Bioinformatics, 2013,29(8):1130.

[10] 李强,万建民. SSRHunter,一个本地化的 SSR 位点搜索软件的开发[J]. 遗传,2005,27(5):808-810.
LI Qiang, WAN Jianmin.SSRHunter: Development of a local searching software for SSR sites[J].Hereditas, 2005, 27(5):808-810.

[11] 李红燕. 基于 BLAST 算法的序列分析软件开发[D]. 长沙:中南大学,2009.
LI Hongyan. Sequence analysis software development based on BLAST algorithm[D].Changsha: Central South University, 2009.

[12] 余劲聪,方柏山. 开发可统计任意密码子用法的软件 BestCodon[J]. 计算机与应用化学,2011,28(1):23-26.
YU Jinchong, FANG Baishan. Development of the BestCodon software for calculating diversified codon usages[J]. Computers and Applied Chemistry,2011,28(01):23-26.

[13] 邱均平, 王菲菲. 基于 SNA 的国内竞争情报领域作者合作关系研究[J]. 图书馆论坛, 2010,30 (6): 34-40.
QIU Junping, WANG Feifei. Analysis on author cooperation relationship of competitive intelligence research in China based on SNA[J]. Library Tribune,2010,30(6): 34-40.

[14] AL-SHAHROUR F, MINGUEZ P, VAQUERIZAS J M, et al. BABELOMICS: a suite of web tools for functional annotation and analysis of groups of genes in high-throughput experiments [J]. Nucleic Acids Research, 2005, 33(suppl 2): W460-W464.

[15] Al-SHAHROUR F, MINGUEZ P, TARRAGA J, et al. FatiGO+: a functional profiling tool for genomic data. Integration of functional annotation, regulatory motifs and interaction data with microarray experiments[J]. Nucleic Acids Research, 2007, 35(suppl 2): W91-W96.

[16] HERRERO J, AL-SHAHROUR F, DIAZ-URIARTE R, et al. GEPAS: A web-based resource for microarray gene expression data analysis [J]. Nucleic Acids Research, 2003, 31(13): 3461-3467.

[17] XIA J, PSYCHOGIOS N, YOUNG N, et al. MetaboAnalyst: a web server for metabolomic data analysis and interpretation[J]. Nucleic Acids Research, 2009, 37(suppl 2): W652-W660.

[18] DONG X, STOTHARD P, FORSYTHE I J, et al. PlasMapper: a web server for drawing and auto-annotating plasmid maps [J]. Nucleic Acids Research, 2004, 32 (suppl 2): W660-W664.

[19] VAN DOMSELAAR G H, STOTHARD P, SHRIVASTAVA S, et al. BASys: a web server for automated bacterial genome annotation [J]. Nucleic Acids Research, 2005, 33(suppl 2): W455-W459.

[20] YE J, FANG L, ZHENG H, et al. WEGO: a web tool for plotting GO annotations [J]. Nucleic Acids Research, 2006, 34(suppl 2): W293-W297.

[21] CHENG J, RANDALL A Z, SWEREDOSKI M J, et al. SCRATCH: a protein structure and structural feature prediction server [J]. Nucleic Acids Research, 2005, 33 (suppl 2): W72-W76.

[22] ZHANG Y. miRU: an automated plant miRNA target prediction server [J]. Nucleic Acids Research, 2005, 33 (suppl 2): W701-W704.

[23] WHITE H D. Pathfindernetworks and author co-citation analysis: A remapping ofparadigmatic information scientists[J]. Journal of the American Society for Information Science and Technology, 2003, 54 (5): 423 -434.

[24] ASHBURNER M, BALL C A, BLAKE J A, et al. Gene Ontology: tool for the unification of biology [J]. Nature Genetics, 2000, 25(1): 25-29.

[25] THOMPSON J D, HIGGINS D G, GIBSON T J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice[J]. Nucleic Acids Research, 1994, 22(22): 4673-4680.

[26] THOMPSON J D, GIBSON T J, PLEWNIAK F, et al. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools[J]. Nucleic Acids Research, 1997, 25 (24): 4876-4882.

[27] HIGGINS D G, THOMPSON J D, GIBSON T J. Using CLUSTAL for multiple sequence alignments[J]. Methods in Enzymology, 1996, 266: 383-402.

[28] 侯剑华,陈悦.战略管理学前沿演进可视化研究[J].科学学研究, 2007, 25(S1): 15-21.
HOU Jianhua, CHEN Yue. Research on visualization of the evolution of strategic management front[J]. Studies in Science of Science,2007, 25(S1): 15-21.

[29] SMITH T F, WATERMAN M S. Identification of common molecular subsequences[J]. Journal of Molecular Biology, 1981, 147(1): 195-197.

[30] KABSCH W, SANDER C. Dictionary of protein secondary structure: pattern recognition of hydrogen - bonded and geometrical features[J]. Biopolymers, 1983, 22(12): 2577-2637.

[31] SAITOU N, NEI M. The neighbor-joining method: a new method for reconstructing phylogenetic trees[J]. Molecular Biology and Evolution, 1987, 4(4): 406-425.

[32] PEARSON W R, LIPMAN D J. Improved tools for biological sequence comparison[J]. Proceedings of the National Academy of Sciences, 1988, 85(8): 2444-2448.

[33] BERMAN H M, WESTBROOK J, FENG Z, et al. The protein data bank[J]. Nucleic Acids Research, 2000, 28 (1): 235-242.

[34] ALTSCHUL S F, MADDEN T L, SCHÄFFER A A, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs[J]. Nucleic Acids Research, 1997, 25(17): 3389-3402.

[35] ASHBUMER M, BALL C A, BLAKE J A, et al. Gene Ontology: tool for the unification of biology [J]. Nature Genetics, 2000, 25(1): 25-29.

[36] ALTSCHUL S F, GISH W, MILLER W, et al. Basic local alignment search tool[J]. Journal of Molecular Biology, 1990, 215(3): 403-410.

[37] BENJAMINI Y, HOCHBERG Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing[J]. Journal of the Royal Statistical Society. Series B (Methodological), 1995,57(1):289-300.

[38] 邱均平, 吕红. 基于知识图谱的国内知识管理发展研

究[J]. 情报学报, 2013, 32(5): 548-560.
QIU Junping, LU Hong. Development research of domestic knowledge management based on mapping knowledge [J].Journal of the China Society for Scientific and Technical Information,2013, 32(5): 548-560.

[39] BELVAUX G, WOLSEY L A. Bc-prod: A specialized branch-and-cut system for lot-sizing problems[J]. Management Science,2000,46(5):724-738.

[40] 张灵芝. 1998 年以来中国高等教育研究热点及其知识可视化图谱分析——基于 CSSCI 高等教育类论文关键词的分析[J]. 高教探索, 2012, (2): 41-46.
ZHANG Lingzhi.Hot issues and knowledge mapping in the research of chinese higher education science 1998[J]. Higher Education Exploration,2012,(2): 41-46.

[41] ROTHER M, ROTHER K, PUTON T, et al. ModeRNA: a tool for comparative modeling of RNA 3D structure[J]. Nucleic Acids Research, 2011, 39(10): 4007-4022.

[42] 欧阳星明, 谢欣荣, 张少伟, 姚小龙. 分子三维结构可视化软件建模的研究与实现[J]. 计算机工程与科学, 2004, 26(5): 25-28.
OUYANG Xingming, XIE Xinrong, ZHANG Shaowei, et al.Research and implementation of the modeling of molecular 3D structure visualization software[J].Computer Engneering & Science,2004, 26(5): 25-28.

[43] DELANO W L. The case for open-source software in drug discovery[J]. Drug Discovery Today, 2005, 10(3): 213-217.

[44] 王非, 郑珩, 杨欣, 张玉彬,吴梧桐. 生物信息学分析软件的开发及在药学领域的应用[J]. 药学进展, 2005, 26(1): 43-48.
WANG Fei, ZHENG Heng, YANG Xin, et al. Development of bio-informatics system and its application in pharmacy[J].Progress in Pharmaceutical Science,2005, 26(1): 43-48.

[45] LANCASTER A K, NUTTER-UPHAM A, LINDQUIST S, et al. PLAAC: a web and command-line application to identify proteins with prion-like amino acid composition [J].Bioinformatics, 2014,30(17):2501-2502.

[46] SEIFERT T, LUND A, KNEISSL B, et al. SKINK: a web server for string kernel based kink prediction in α-helices[J].Bioinformatics, 2014,30(12):1769-1770.